Cultural Cuisine Analysis

Ido Shakartsi - ID: 207902834 - CSE: idoshakar
Ohad Persky - ID: 208198804 - CSE: ohad.persky
Michael Terkeltoub - ID: 322695107 - CSE: p820395

Problem Description

Despite the abundance of online recipes, there's still no good way to explore and compare ingredient patterns across different cuisines. What ingredients make a cuisine unique? Are there hidden similarities between foods from totally different parts of the world? And how does nutrition vary between these culinary traditions?

This project is about digging into recipe data from various cultures to find patterns, trends, and nutritional differences. By analyzing the ingredients used in different cuisines, we want to build a better understanding of what defines them, how they relate to each other, and how we might make healthier food choices without losing the essence of a dish.

As part of this, we also plan to build a recommendation system or prediction model - for example, to suggest healthier ingredient swaps.

Data

1. Food Nutrition Dataset (Kaggle)

Link: https://www.kaggle.com/datasets/utsavdey1410/food-nutrition-dataset
Content: Core nutritional values for individual ingredients
Data Size: 1.13MB.

2 . USDA FoodData Central

Link: https://www.kaggle.com/datasets/thedevastator/the-nutritional-content-of-food-a-comprehensive
Content: More detailed nutritional values, extending the previous dataset.
Data Size: 200

4. What's Cooking Recipes Dataset (Kaggle)

Link: https://www.kaggle.com/c/whats-cooking/data
Content: Main dataset with **20,000+ recipes** spanning **20+ cuisines**.
Data Size: 12MB

Your Solution

We categorize cuisines and their recipes along key cultural and ingredient-based dimensions:

- **Core Ingredients** – Staples, proteins, flavor bases, and oils/fats that define each cuisine.
- **Ingredient Diversity & Frequency** – Number of unique ingredients and dominant combinations, revealing patterns and similarities.

- **Geography** – Grouping cuisines by environment to understand how natural conditions shape food traditions.
- **Food Groups** – Distribution of ingredients into broader categories (like herbs/spices, proteins, vegetables, etc.) to see the balance and culinary “signature” of each cuisine.

“What’s cooking?” is an online database from Kaggle.com containing 39,774 recipes in a json file from 20 different cuisines, weighing ~12 MB’s.

In our mission to try and compare and analyze differences and similarities across recipes, we needed to normalize the ingredient field. We achieved that by finding an ingredient space that we will normalize each ingredient to.

We went through three different datasets of ingredients throughout the development process. The first one “food ingredients” from Kaggle.com was small and did not cover a large amount of ingredients. The second one was foodb.ca. We proceeded to normalize the data, by searching for each ingredient in the recipe in google coupled with the term 'food ca'. After multiple iterations, using SerpAPI, and a Google custom search engine, that proved to be inadequate, we began exploring other options. The one that proved the most viable was a local search engine on top of DuckDB. It was also an effective solution to the RAM usage problem, and it proved itself in subsequent stages of the project. To handle ingredients that couldn't be matched through the database search, we implemented a fuzzy matching system using RapidFuzz with hierarchical scoring. This approach uses approximate string similarity based on Levenshtein distance, enhanced with word overlap analysis. The system applies different scoring weights depending on the number of common words between ingredients (≥3 words: 0.6 base + adjustments, 2 words: 0.4 base + adjustments, 1 word: 0.2 base + adjustments), with additional bonuses for exact subset matches. For performance, results are cached and limited to the top 5000 candidates. This fuzzy fallback proved essential for handling ingredient variants, misspellings, and alternative phrasings not present in the structured database.

We noticed that while foodb.ca was very concise, it was also lacking in scale. Data normalization showed that there were still a lot of ingredients missing, and the cuisine-ingredient heatmap did not give us the unequivocal results we were expecting for some of the cuisines. This prompted us to look for other datasets. The one with the most foods was https://fdc.nal.usda.gov/ - USDA FoodData Central. It had about 2 million entries, and its size was about 200 MB. However, while having the most variety of foods, it also contained a major amount of duplication. Later on we would have to cluster recipes by their ingredients, and that would not work with similar ingredients being classified as different from each other. So our solution was to cluster by a combination of the most frequent nutritional values, and the encoding of the ingredients names. Using a NN model to encode the names of the ingredients was too slow, and not very beneficial, since we were not encoding sentences. Therefore we decided to use GloVe, which disregards each word's role in the sentence. We also had a lot of trouble regarding RAM usage. The solution was to use memory maps for every stage of the encoding process. We began by using batched Birch clustering, however it was very slow, and some of the clusters were very large and contained unrelated entries. We wanted spherical clustering, which was very convenient, since fast methods often produce these results (there is

no need for structure or connectivity). We briefly considered X-Means, however due to its nature it did not allow for batch processing, since it measures the distance between every set of points. We also did not want a predetermined amount of clusters, so we chose Bisecting Mini Batch K Means, where at each iteration, all clusters are divided in two. It was blazing fast, and produced very adequate results. The resulting dataset, where each food was assigned a cluster, was plugged into the search pipeline.

After this stage, we had the What's cooking recipe database with normalized ingredients, and we proceeded to the next stage - frequent itemsets.

The frequent itemset mining stage applies the A-priori algorithm to discover patterns of ingredient clusters that commonly appear together across recipes. This unsupervised pattern discovery serves as the foundation for the subsequent clustering analysis, transforming the normalized ingredient data into meaningful co-occurrence features.

The A-priori algorithm operates through iterative passes, beginning with frequent 1-itemsets (individual ingredient clusters that appear in at least a threshold number of recipes) and progressively building larger itemsets. The algorithm leverages the downward closure property: if an itemset is frequent, all its subsets must also be frequent. This enables aggressive pruning of candidate itemsets, significantly reducing computational complexity from exponential to manageable levels.

Support Threshold Selection (30 recipes): The support threshold of 30 was chosen to balance pattern significance with computational tractability. With 39,774 total recipes, this represents approximately 0.08% support, filtering out extremely rare ingredient combinations while preserving meaningful patterns. This threshold filters out statistical noise, and helps in ensuring that discovered itemsets represent genuine culinary relationships, while maintaining sufficient itemsets for meaningful clustering analysis. The threshold also accounts for the fact that ingredient clusters already provide abstraction - individual clusters may represent multiple related ingredients, so co-occurrence patterns at the cluster level should be reasonably frequent to be culinarily meaningful.

Itemset Size Progression (2-5 itemsets): The analysis examines itemsets of sizes 2 through 5, with larger sizes terminated when no frequent patterns emerge. This range captures the most relevant culinary combinations - pairs representing basic flavor pairings, triplets capturing foundational recipe structures, and 4-5 itemsets representing more complex recipe signatures. The selection of itemsets in the 2-5 range was empirically validated through comprehensive frequent itemset analysis examining support decay patterns, coverage efficiency, and data availability across different itemset sizes.
Support values follow an exponential decay pattern, dropping from ~0.01 for 1-itemsets to ~0.001 for larger itemsets. The sharp decline after size 3 indicates that itemsets larger than 5 contain insufficient support to reliably represent meaningful culinary patterns, providing theoretical justification for focusing on the 2-5 itemset range where support remains above the noise threshold.

This analysis measures the practical utility of different itemset sizes by calculating average support among the top 200 itemsets for each size. The results show 1-2 itemsets achieve maximum efficiency (~0.04 average support), with performance dropping dramatically after size 3. This empirical evidence demonstrates that selecting itemsets in the 2-5 range maximizes the discriminative power per feature, directly validating the methodological choice to avoid larger itemsets that provide diminishing returns

This plot confirms adequate data volume exists to support the analysis, showing peak availability of itemsets at sizes 3-4 (approximately 30,000 itemsets each) before declining sharply for larger sizes. The substantial quantity of available 2-5 itemsets ensures that the selection strategy is not artificially constrained by data scarcity, while the dramatic drop-off for sizes 6+ indicates these larger itemsets are genuinely rare patterns rather than systematically discoverable relationships.

The RecSystem We decided to build a simple recommendation system, which receives a number of ingredients (up to six), and searches for an appropriate recipe. The solution was quite straightforward - look for the cluster with the highest Jaccard similarity to the given ingredient list (common frequent itemsets), and return all recipes within that cluster which had the most common with the ingredient list.

The prediction model We wanted to show that there is high correlation between the ingredients used in the recipes, and the cuisines to which they belong to. This prompted us to use a neural network to predict the cuisine of a recipe. We used a simple two layer network with BCE loss (a recipe was allowed to be classified as belonging to only one cuisine, CE loss gave us very bad results), which achieved 0.08 loss.

Analysis Structure

- **Per-Cuisine Analysis**: Each cuisine is examined with word clouds of most common ingredients as well as least common ingredients (excluding outliers), number of unique ingredients, recipe length distribution, top 10 N-ingredient combinations, recipe similarity (Jaccard and cosine similarity), major food groups, most common ingredient in each group, and an ingredient rarity index.

- **Cross-Cuisine Comparison**: We compare cuisines with heatmaps of top ingredients, recipe complexity, recipe similarity by geography, frequent item sets across cuisines of different sizes, pairs of most similar cuisines by ingredients, and stacked bar plots of food group distributions.

This approach uncovers defining patterns for each cuisine, highlights cross-cultural similarities, and shows how ingredient groups shape culinary identity across regions.

Global Cuisine Analysis: Main Findings

Overview

We looked at ingredient patterns, diversity, and similarities across 20 different cuisines. The results give some interesting points about how food traditions connect and differ around the world.

Similarities Between Cuisines

One of the strongest results is the high similarity between Italian and French (0.82). Italian also connects a lot with Mexican (0.79) and Southern US (0.79). That shows some cooking basics are shared even if the cuisines come from different places.
In Europe, the Mediterranean group (Italian, French, Spanish, Greek) clearly clusters together. The Southern US also shows up as a kind of “bridge,” linking European and Latin American styles.
In Asia, there’s a clear regional block. Thai and Vietnamese (0.76) are very close, as well as Chinese and Thai (0.75). The similarity heatmap makes these clusters really obvious, with warmer colors grouped by region.

Ingredient Use and Diversity

Most cuisines use about 10–12 ingredients per recipe, but some vary more. Moroccan, Vietnamese, and Cajun/Creole came out with the highest diversity scores, which makes sense given their heavy spice use and mixing of traditions. Irish and Brazilian were at the lower end, meaning the recipes rely on fewer types of ingredients.

Food Group Patterns

Looking at food groups, herbs and spices are dominant almost everywhere (20–45%). Moroccan and Indian are especially spice-heavy. Asian cuisines (Chinese, Japanese, Korean) show higher use of soy products and seafood, while European cuisines spread more evenly across food groups. Mexican and Jamaican stand out with strong spice blends and peppers.

Common vs. Regional Ingredients

Some ingredients are almost universal. Salt shows up in most recipes (60–77%), along with onions and garlic. But each region has its own signature items: soy sauce and ginger in Asian cuisines, olive oil and herbs in Mediterranean, and chili peppers and spice mixes in Latin American.

Takeaway

The big picture is that cuisines keep their unique character through certain ingredients and techniques, but at the same time, you can see clear connections across continents. High diversity in cuisines like Moroccan and Vietnamese ties back to trade routes and cultural mixing. And the Italian–Mexican similarity shows that food traditions can overlap even when the regions are far apart.